Intro to R for Epidemiologists

Lab 7 (2/26/15)

Data

This lab will use the `hflights` dataset within the `hflights` R package. Recall that you need to load the hflights library (`library(hflights)` before you can access the data. This dataset contains information on flights departing from Houston's airports in 2011 (Source: Bureau of Transporation, Research and Innovation Technology Administration).

Part 1. Simple Linear Regression

1. Subset your dataset to only include observations in March.
2. Remove any missing data from your dataset.
3. Look at histograms and scatterplots of flight time and distance. Are the linear regression assumptions met?
4. Fit a simple linear model to assess whether distance is associated with flight time. Use (natural) log distance and log flight time and interpret the results.
5. What is the R^2 of your model?

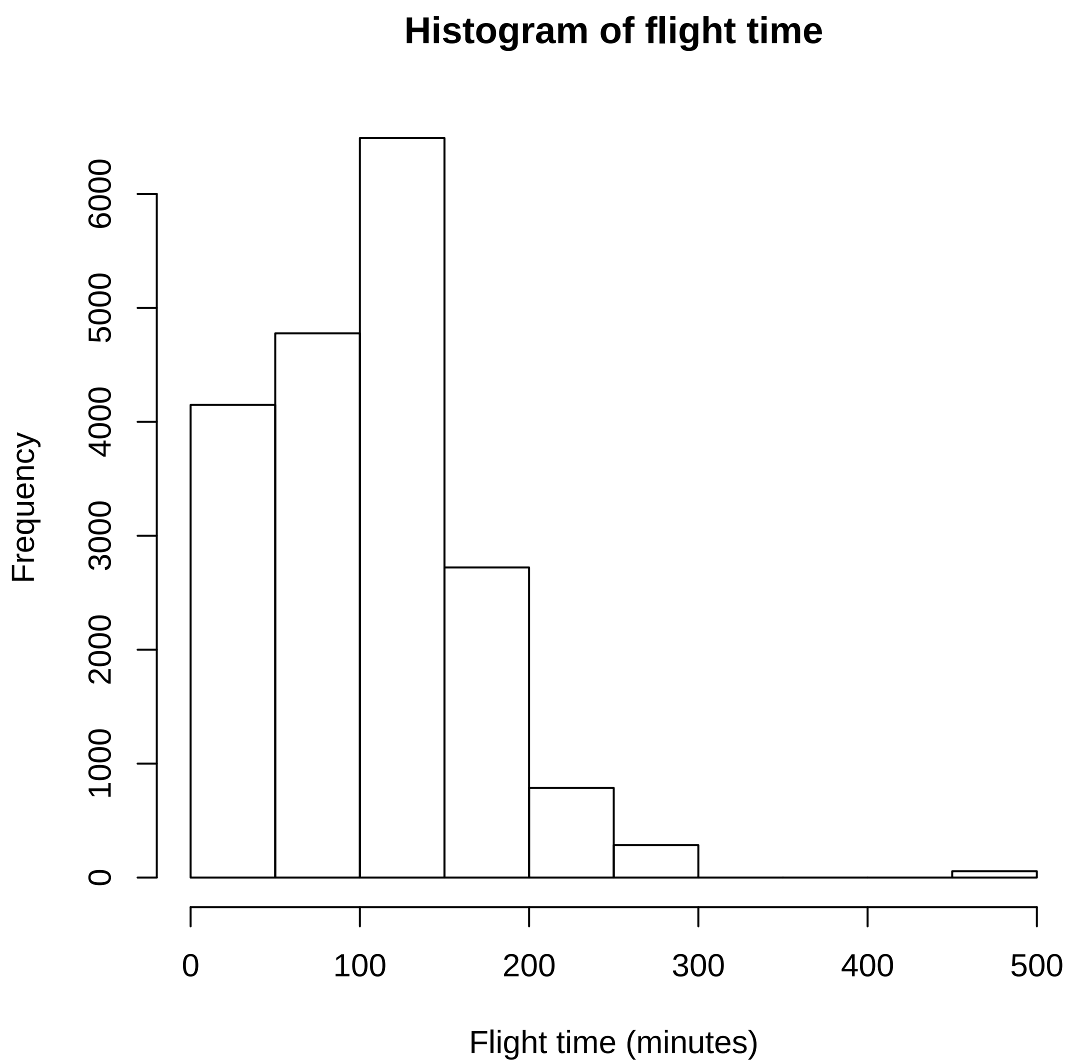
```
# Part 1
library(dplyr)
library(hflights)
data(hflights)

# 1 Restrict to March
march <- filter(hflights, Month == 3)

# 2 Look at summary statistics
summary(hflights$AirTime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    11.0    58.0   107.0   108.1   141.0   549.0    3622
```

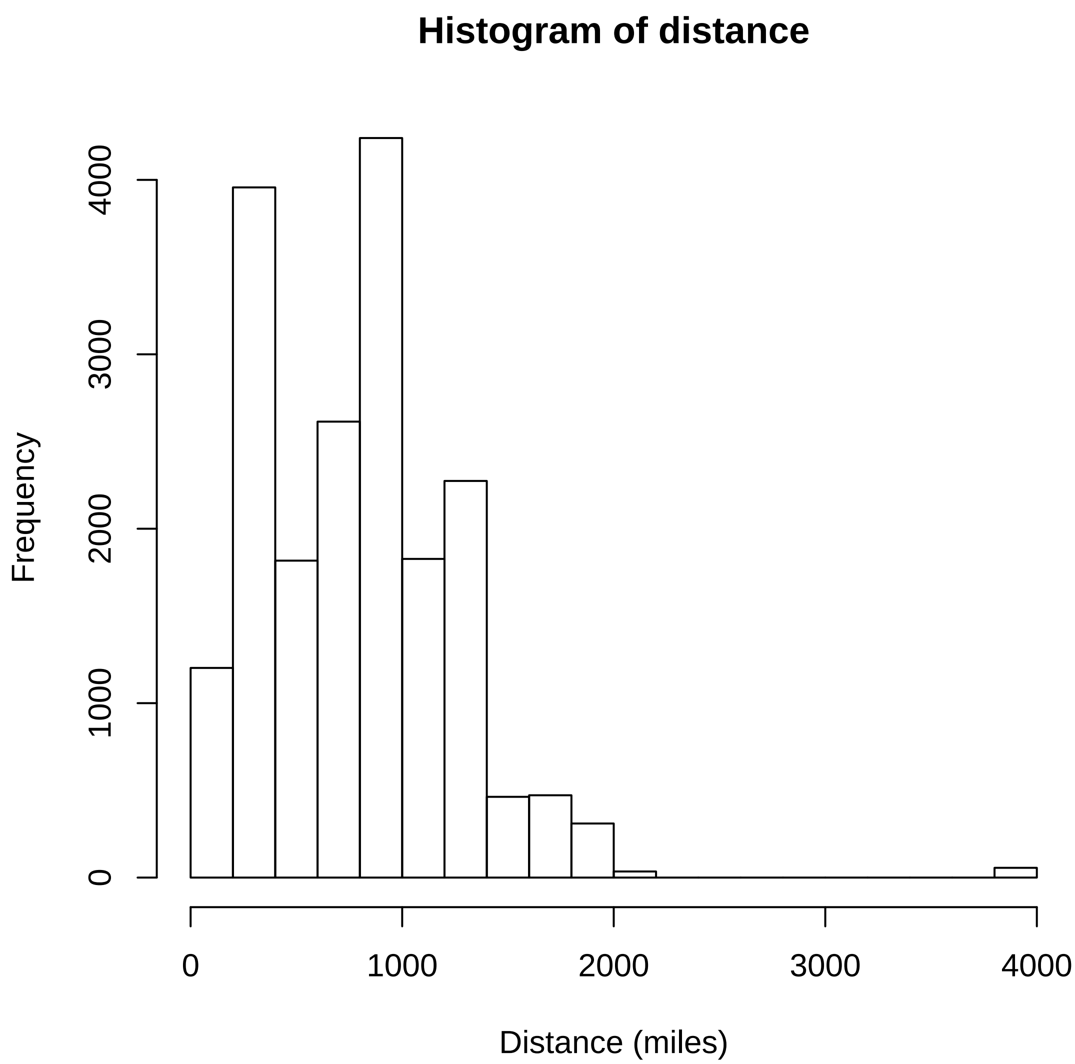
```
summary(hflights$Distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    79.0   376.0   809.0   787.8  1042.0  3904.0
```

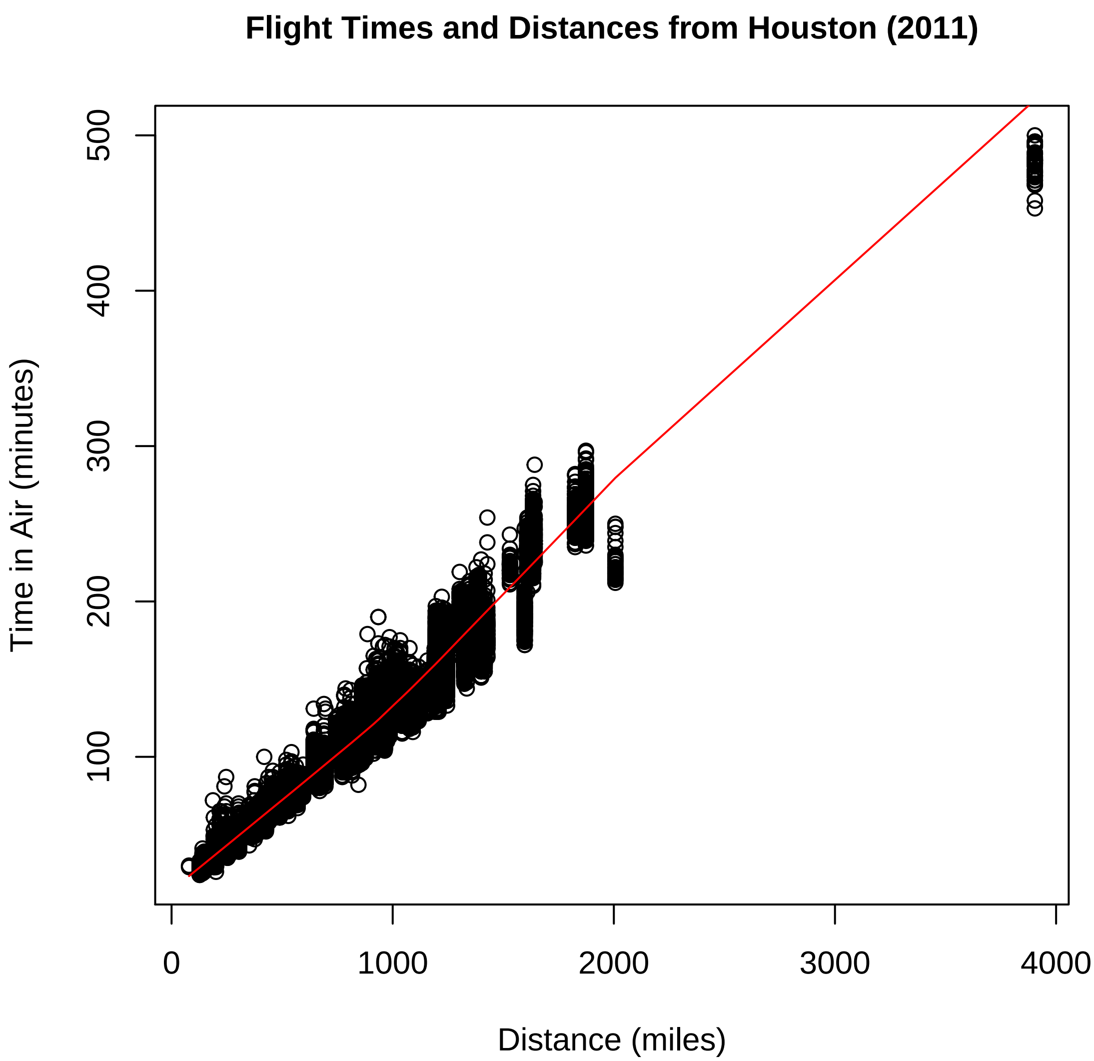
```
# Complete case data
march_cc <- march[complete.cases(march), ]

# 3 Histograms of airtime and distance
hist(march_cc$AirTime, main = "Histogram of flight time", xlab = "Flight time (minutes)")
```



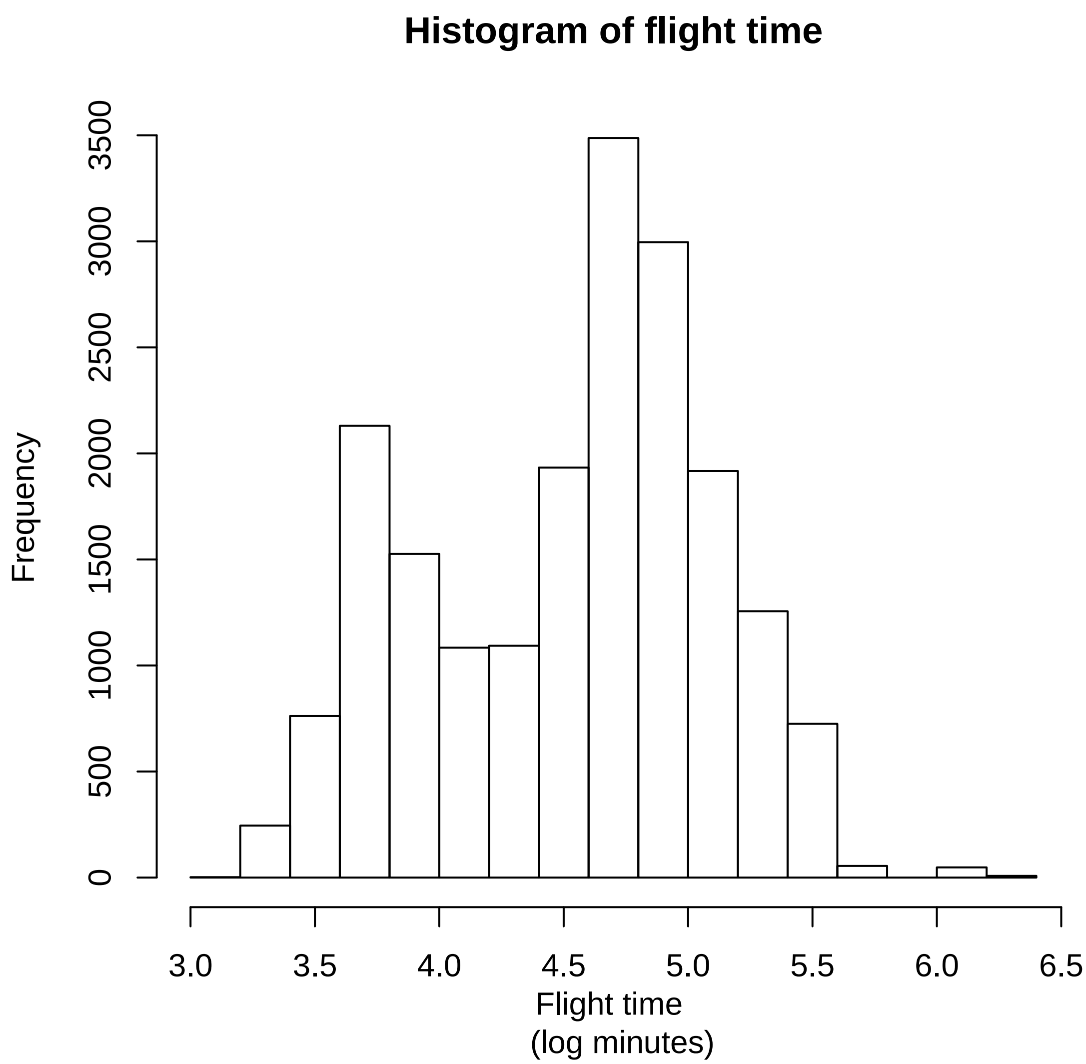

```
hist(march_cc$Distance, main = "Histogram of distance", xlab = "Distance (miles)")
```



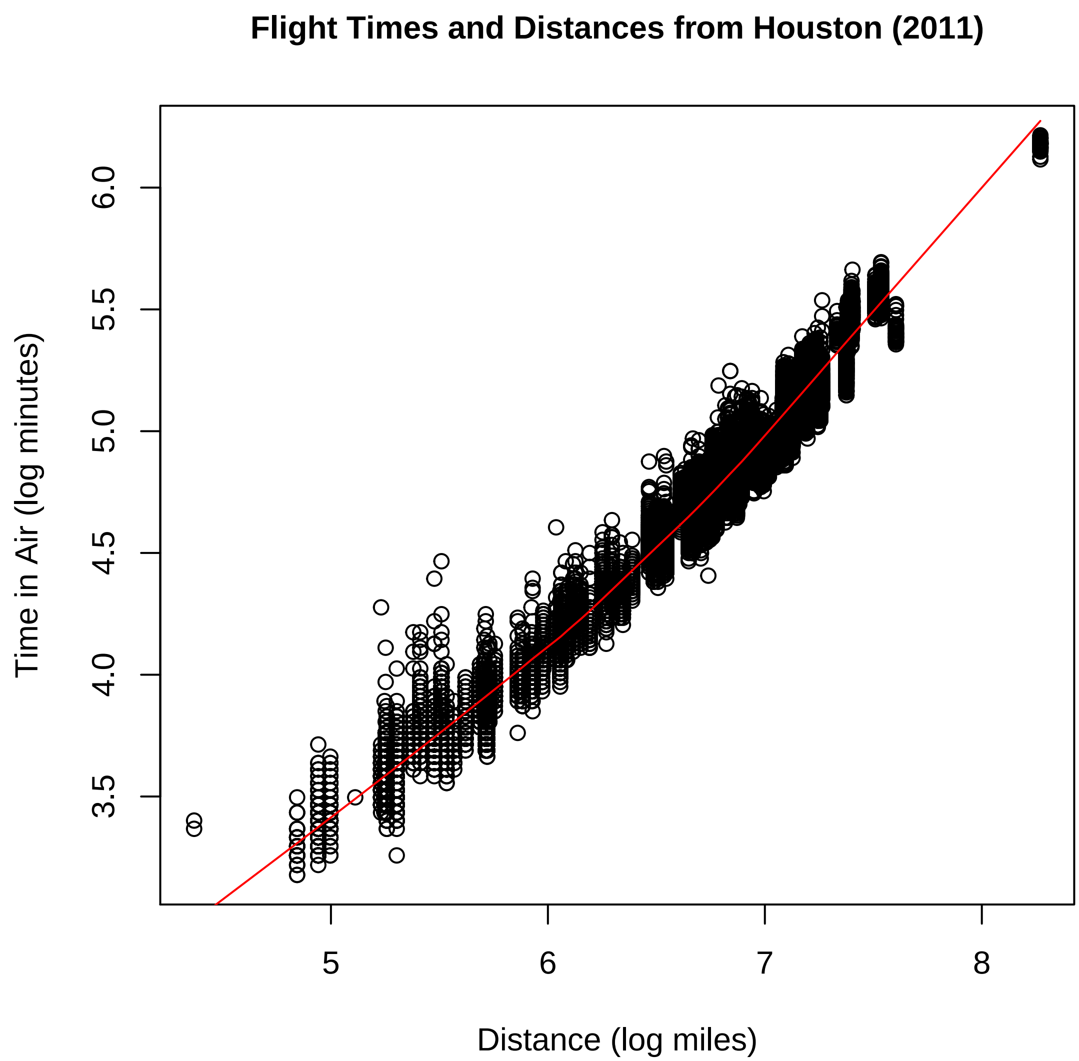

```r
# Scatterplot
low1 <- lowess(x = march_cc$Distance, y = march_cc$AirTime)
plot(march_cc$Distance, march_cc$AirTime, xlab = "Distance (miles)", ylab = "Time in Air (minutes)",
    main = "Flight Times and Distances from Houston (2011)", cex.lab = 1, cex.axis = 1,
    cex.main = 1)
lines(low1$x, low1$y, col = "red")
```




```
# Create logged variables
march_cc <- mutate(march_cc, lAirTime = log(AirTime), lDistance = log(Distance))
# Look at histogram of air time
hist(march_cc$lAirTime, main = "Histogram of flight time", xlab = "Flight time \n  (log minutes)")
```




```
# look at scatterplot
low1 <- lowess(x = march_cc$lDistance, y = march_cc$lAirTime)
plot(march_cc$lDistance, march_cc$lAirTime, xlab = "Distance (log miles)", ylab = "Time in Air (log minu
    main = "Flight Times and Distances from Houston (2011)", cex.lab = 1, cex.axis = 1,
    cex.main = 1)
lines(low1$x, low1$y, col = "red")
```




```r
# The normality assumption of air time was not met so we logged air time

# The linearity assumption of air time vs. distance was not met, so we
# logged distance

# The equal variances assumption seems to be met

# There may be some outlying values with large distances that we could
# consider removing from the analysis


# 4 Fit a simple linear regression model
fit1 <- lm(lAirTime ~ lDistance, data = march_cc)
sfit1 <- summary(fit1)
sfit1$coef
```

```
##               Estimate  Std. Error   t value Pr(>|t|)
## (Intercept) -0.7881721 0.006998809 -112.6152        0
## lDistance    0.8230587 0.001076689  764.4347        0
```

```r
# Log distance and log air time are associated for Houston flights in March
# 2011.

# 5 Extract out r.squared
sfit1$r.squared
```

```
## [1] 0.9680845
```

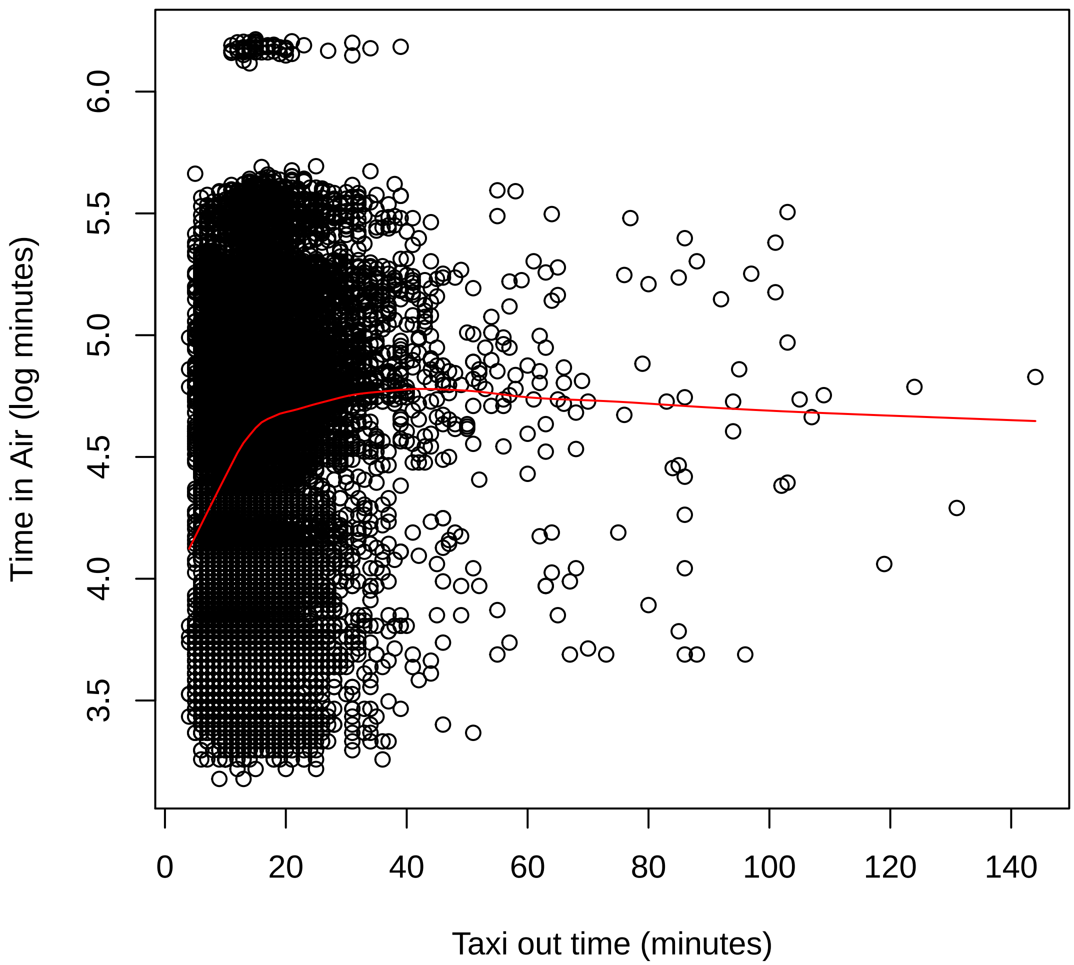
```r
# log distance explains 96.8% of the variability in log air time
```

Part 2. Multiple Linear Regression

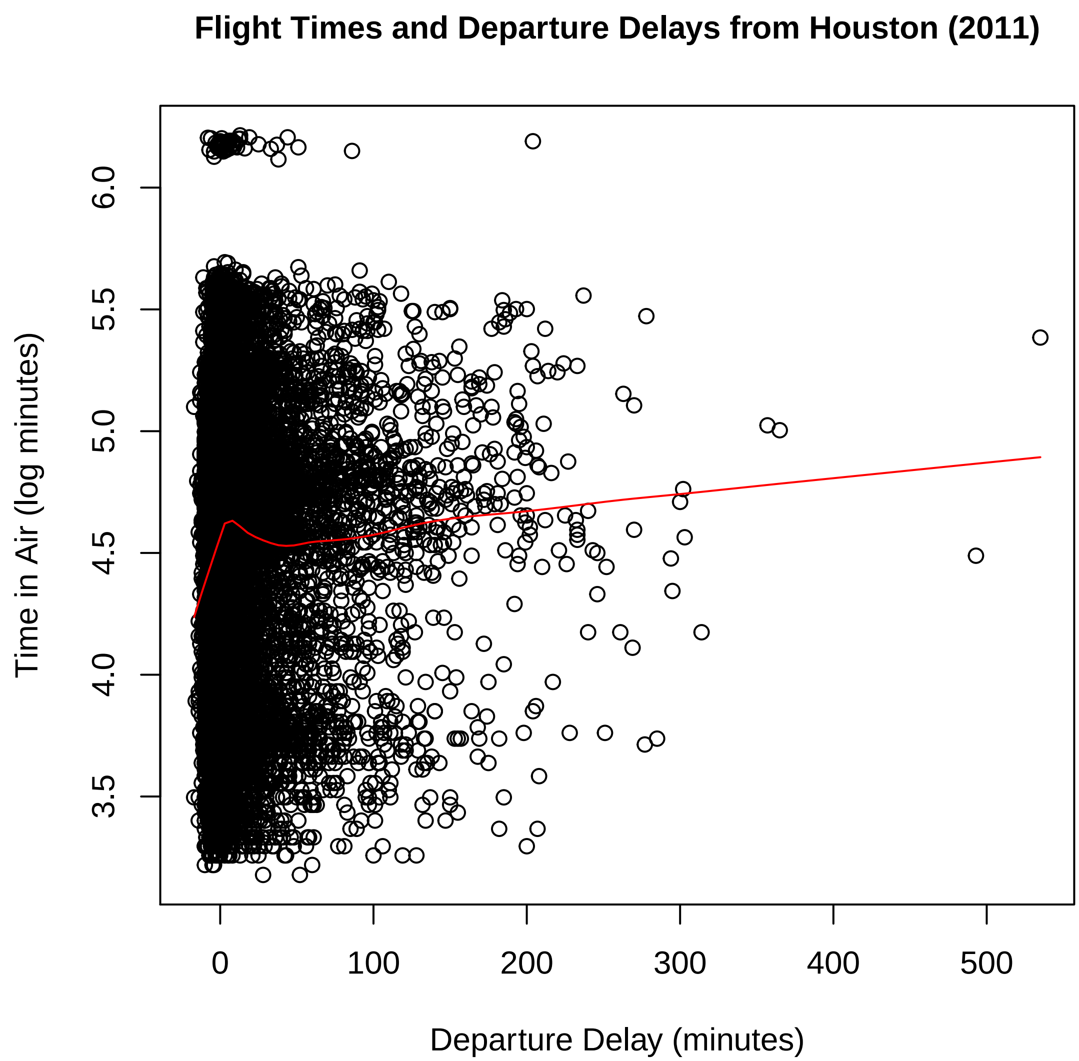
1. Use scatterplots to explore the associations between log flight time and taxi out time and log flight time and departure delay.
2. Fit a multiple linear regression model to determine whether log distance, log taxi out time, and departure delay (not logged) are associated with log flight time.
3. Create a vector of the estimated coefficients from your model in (2).
4. Find 95% confidence intervals for these coefficients (Hint: `?confint`).
5. Does departure delay or taxi out time confound the association between distance and air time?

```r
# Part 2 1 Plot scatteplots of data
low1 <- lowess(x = march_cc$TaxiOut, y = march_cc$lAirTime)
plot(march_cc$TaxiOut, march_cc$lAirTime, xlab = "Taxi out time (minutes)",
    ylab = "Time in Air (log minutes)", main = "Flight Times and Taxi out times from Houston (2011)",
    cex.lab = 1, cex.axis = 1, cex.main = 1)
lines(low1$x, low1$y, col = "red")
```

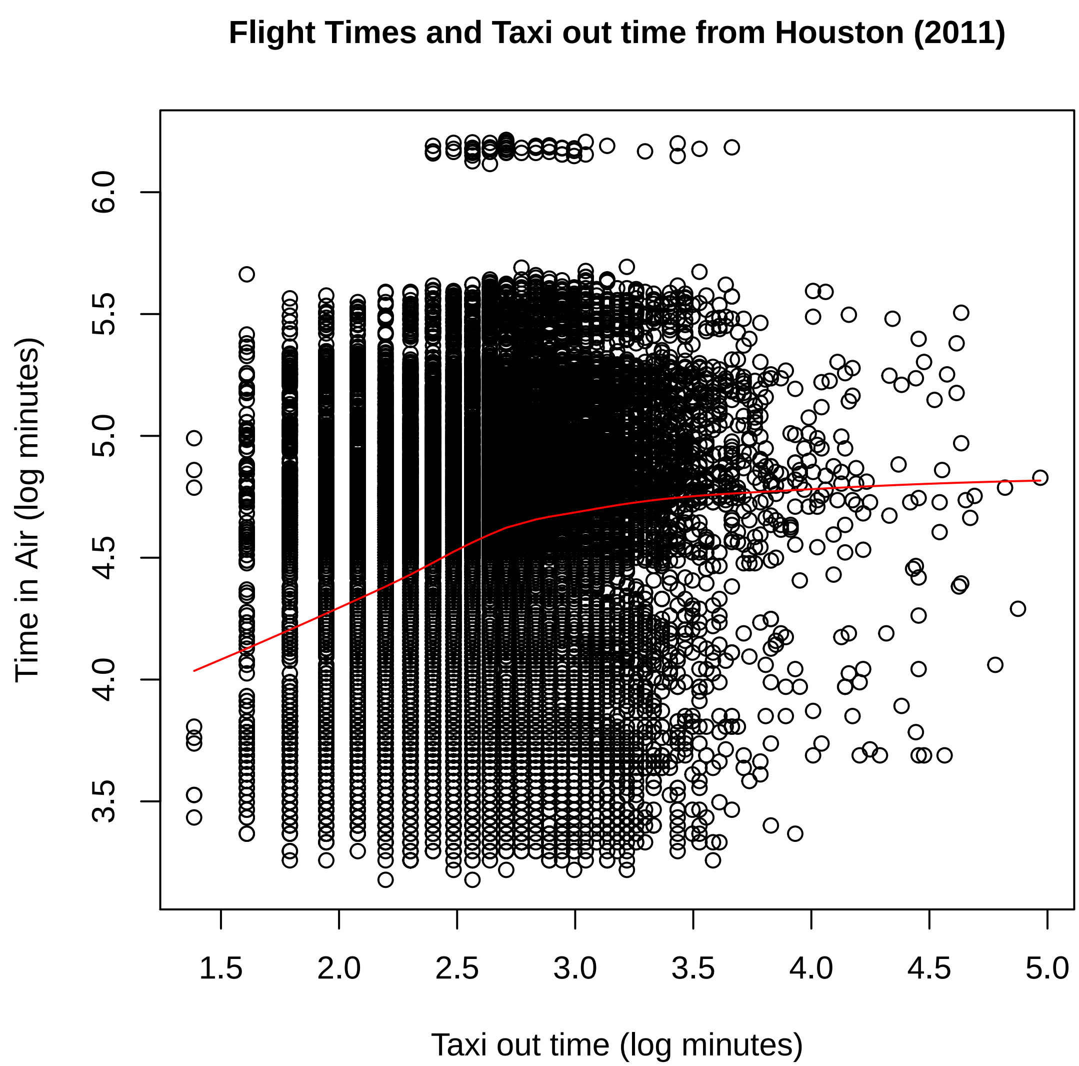
Flight Times and Taxi out times from Houston (2011)




```r
low1 <- lowess(x = march_cc$DepDelay, y = march_cc$lAirTime)
plot(march_cc$DepDelay, march_cc$lAirTime, xlab = "Departure Delay (minutes)",
    ylab = "Time in Air (log minutes)", main = "Flight Times and Departure Delays from Houston (2011)",
    cex.lab = 1, cex.axis = 1, cex.main = 1)
lines(low1$x, low1$y, col = "red")
```




```r
# Create logged taxi out time
march_cc <- mutate(march_cc, lTaxiOut = log(TaxiOut))

# Plot logged taxi out time and logged air time
low1 <- lowess(x = march_cc$lTaxiOut, y = march_cc$lAirTime)
plot(march_cc$lTaxiOut, march_cc$lAirTime, xlab = "Taxi out time (log minutes)",
    ylab = "Time in Air (log minutes)", main = "Flight Times and Taxi out time from Houston (2011)",
    cex.lab = 1, cex.axis = 1, cex.main = 1)
lines(low1$x, low1$y, col = "red")
```




```r
# Fit regression model
fit2 <- lm(lAirTime ~ lDistance + lTaxiOut + DepDelay, data = march_cc)
sfit2 <- summary(fit2)
sfit2$coef
```

```
##                  Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept) -8.149529e-01 7.560935e-03 -107.784670 0.000000e+00
## lDistance    8.209224e-01 1.101589e-03  745.216416 0.000000e+00
## lTaxiOut     1.568672e-02 1.698018e-03    9.238250 2.759327e-20
## DepDelay    -5.418667e-05 2.564984e-05   -2.112554 3.465181e-02
```

```r
# 2 Extract out coefficients
beta_hat <- sfit2$coef[, 1]

# 3 Get confidence intervals
confint(fit2)
```

```
##                     2.5 %        97.5 %
## (Intercept) -0.8297729588 -8.001328e-01
```

```
## lDistance    0.8187631746  8.230816e-01
## lTaxiOut     0.0123584544  1.901498e-02
## DepDelay    -0.0001044626 -3.910745e-06
```

Part 3. ANOVA

1. Perform an ANOVA examining the relationship between log(AirTime) and destination. What hypothesis does this test? What is your conclusion?
2. Extract the F-statistic from the linear model (`lm` object) and the F-statistic from the ANOVA.

```r
# Part 3 1 Fit anova
fit3 <- lm(lAirTime ~ Dest, data = march_cc)
anova_air <- anova(fit3)

# Extract F statistic from ANOVA
anova_air$F[1]
```

```
## [1] 16361.99
```

```r
# Extract F statistic from linear model
summary(fit3)$fstatistic[1]
```

```
##    value
## 16361.99
```